

CMS

CERN

LHC

[ Large Hadron Collider ]

# Analysis Preservation Open Data, RECAST

ALICE

L Heinrich  
Snowmass 2020

LHCb

ATLAS

CMS

CERN

**Lest we forget: the LHC is special**

LHC

[ Large Hadron Collider ]

- **The data and results extracted from it are unique**
- **The analyses used to extract result from data are also unique**

**What is our scientific output beyond the papers?**

- **How do we make our results most useful?**
- **What data can we make public and in what formats?**
- **How can we - as Collaborations - exploit the analyses we have invested in?**

ALICE

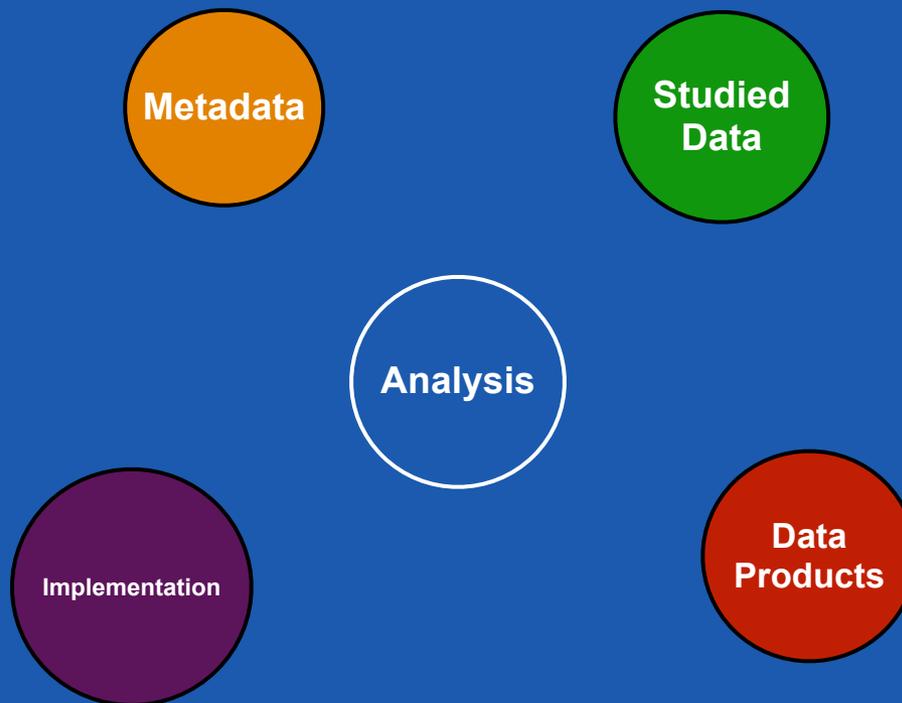
LHCb

ATLAS

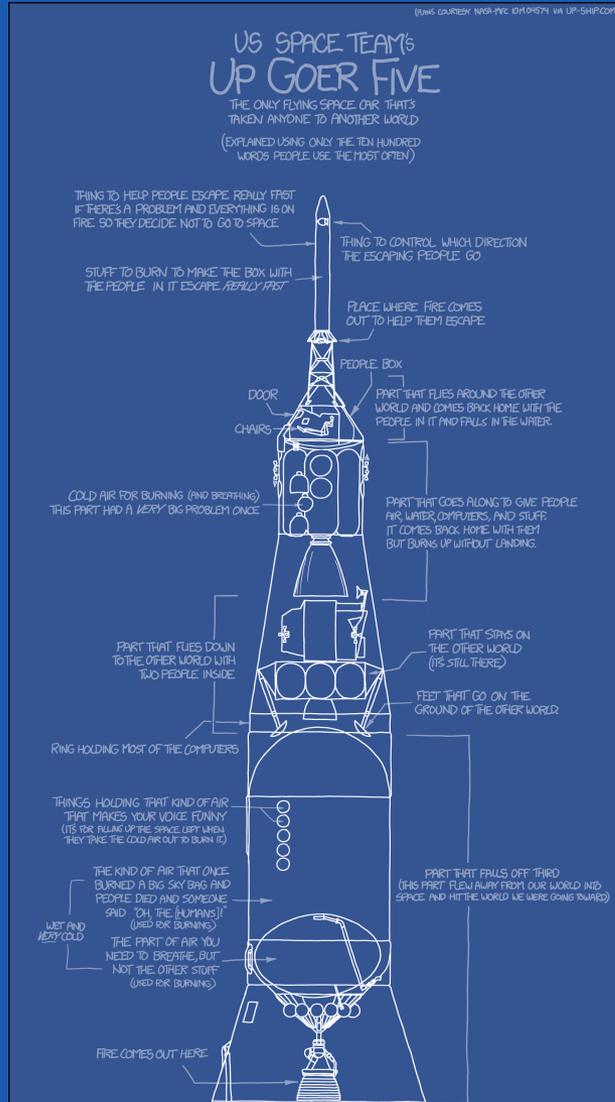
# Analysis Sketch



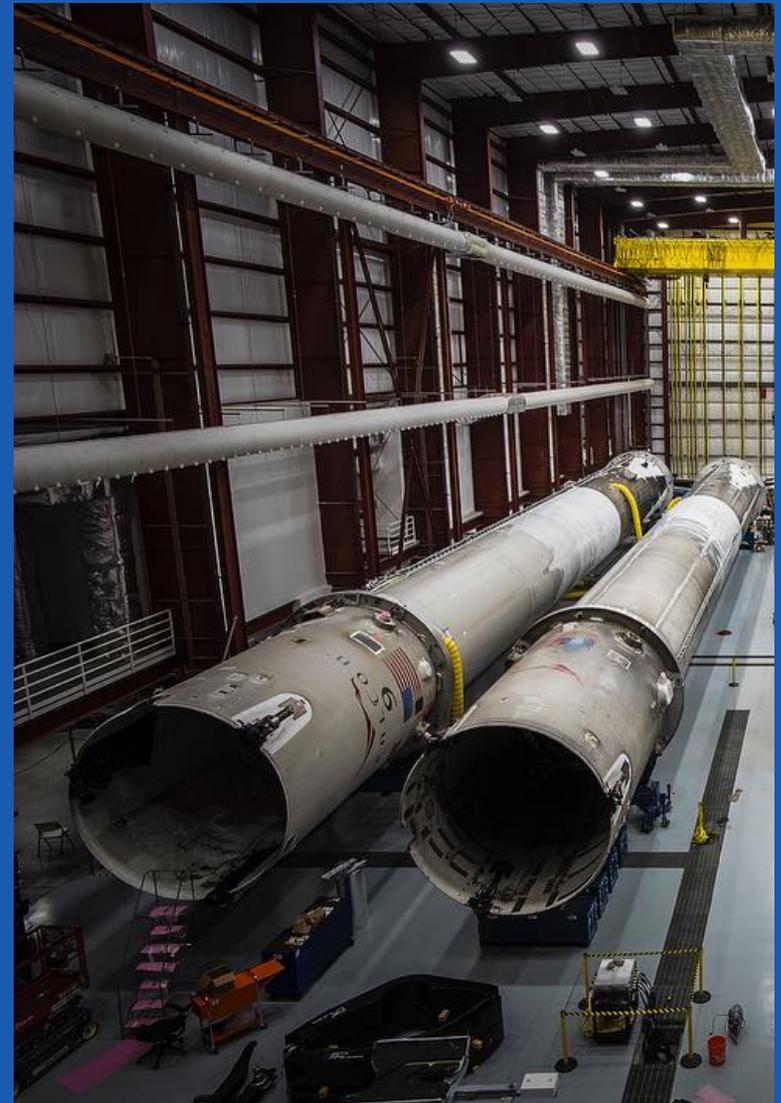
# Preservation Domains



# External

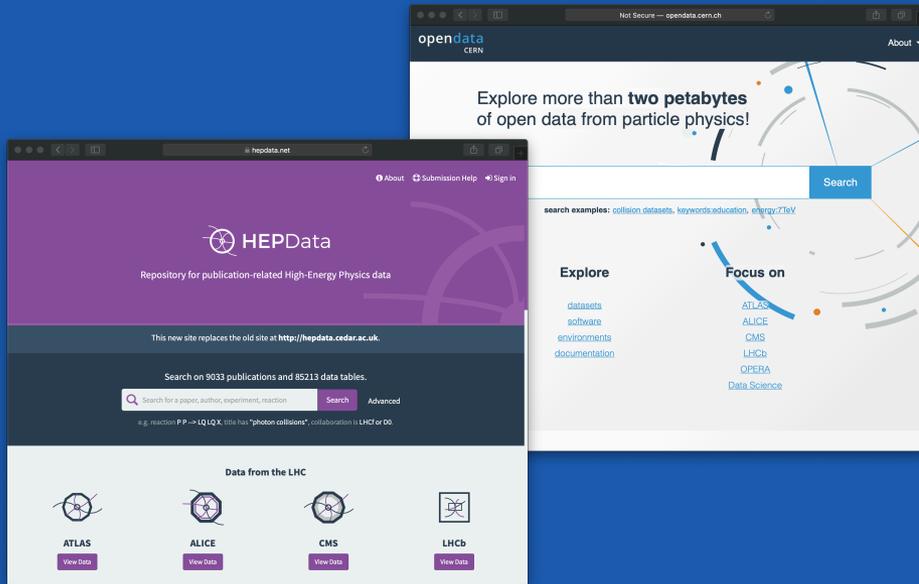


# Internal



# Three broad areas of activity

## External



Analysis Data Products  
and Result Preservation

Open Data for Outreach,  
Education and Research

## Internal



Reproducible Workflows &  
Analysis Preservation



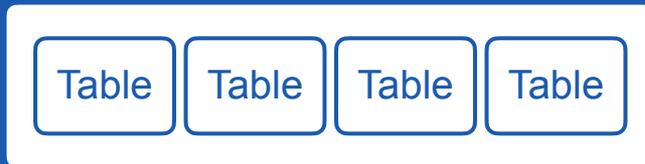
HepData has been the main vehicle to provide

high quality public **data products** for published analyses

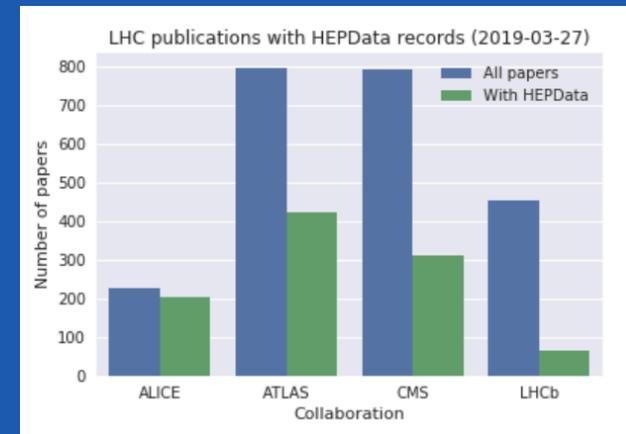
**publicly available**. All LHC experiments rely on this.

- HepData submission often required for analysis approval

Types of data products expanded from



to broader collection of data



ALICE: 90%  
ATLAS: 52%  
CMS: 39%  
LHCb: 14%

# Additional Material helps approximate reimplementa- tion of data analyses w/ e.g. Rivet (can cover also BSM and HI)

```
#include "SimpleAnalysis/AnalysisClass.h"
#include "SimpleAnalysis/NtupleMaker.h"
#include "SimpleAnalysis/PDFRweight.h"
#include <LHAPDF/LHAPDF.h>
#include "TMath.h"

DefineAnalysis(EwkOneLeptonTwoBjets2018)
// Wh->l+l+bb+met analysis (Run2 data)

void EwkOneLeptonTwoBjets2018::Init()
{
    // Define signal/control regions
    // ...
}
```

```
// -*- C++ -*-
#include "Rivet/Analysis.h"
#include "Rivet/Projections/ChargedFinalState.h"
#include "Rivet/Tools/Correlators.h"
#include "Rivet/Tools/AliceCommon.h"
#include "Rivet/Projections/AliceCommon.h"

namespace Rivet {

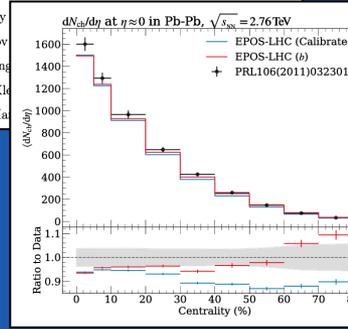
    /// @brief Multiparticle azimuthal correlations pp, pPb, XeXe and Pb
    class ALICE_2019_I1723697 : public CumulantAnalysis {
    public:

        /// Constructor
        ALICE_2019_I1723697() :
            CumulantAnalysis("ALICE_2019_I1723697") {}
    };
};
```

## Confronting Experimental Data with Heavy-Ion Models

Christian Bierlich,<sup>1,2</sup> Andy  
Peter Harald Lindeno  
Jan Fiete Grosse-Oetring  
Patrick Kirchgaesser,<sup>6</sup> Jochen Kl  
Christine O. Rasmussen,<sup>2</sup> Ma

### RIVET for Heavy Ions



C++ Code Snippets as starting point, or (better) full Rivet Routine

Efficiency Maps:

ML models uploaded to HepData

[ATLAS Record]

**Figure 3 (2D effs)** <https://www.hepdata.net/record>

Contents of Figure 3.

The product of the acceptance and efficiency in the  $\epsilon\tau_0$  vs.  $m_g$  plane for the GMSB model, after all requirements.

**cmenergies** 13000.0

**observables** ACC

**phrases** Dark Matter, GMSB, Long-Lived, Jet Production

**reactions** P P -> X X

**xml File**

Final BDT weights for BDT\_high for 'even' events, ROOT 6.04/16; TMVA 4.2.1

**Download**

**Variables**

**Trees**

[CMS Record]





**HepData is for experimentalists interacting with wider community by releasing public information about already existing data analyses.**

**But it **may not be enough** to interact with community on developments of new analyses techniques.**

**For this we might require a more free-form mode of collaboration: Open Data**

# Open Data

## All LHC Experiments have Open Data Programs

- integrated into CERN Open Data Portal
- ATLAS, LHCb, ALICE so far focused mainly on Outreach & Education

Event Display Exercise

Information of selected particle

Particles saved K<sup>+</sup> and pi<sup>+</sup>

Close up on collision

Projections

Detector opacity

Event display

$W \rightarrow \ell \nu$

$Z \rightarrow \ell \ell$

$t\bar{t} \rightarrow \ell \nu b q \bar{q} \bar{b}$

Transverse Mass of the W Candidate

ATLAS Open Data  
 $\sqrt{s} = 13 \text{ TeV}$   
 $\int L dt = 3.2 \text{ fb}^{-1}$   
 $W \rightarrow \ell \nu$

Invariant Mass of the Z Candidate

ATLAS Open Data  
 $\sqrt{s} = 13 \text{ TeV}$   
 $\int L dt = 3.2 \text{ fb}^{-1}$   
 $Z \rightarrow \ell \ell$

Lepton Transverse Momentum

ATLAS Open Data  
 $\sqrt{s} = 13 \text{ TeV}$   
 $\int L dt = 3.2 \text{ fb}^{-1}$   
 $t\bar{t} \rightarrow \ell \nu b q \bar{q} \bar{b}$

D\* Lifetime Exercise

Step 7: compare results to the PDG value

ALICE analysis modules

WELCOME to ROOT

Version 5.34/08 31 May 2013

You are welcome to visit our Web site  
<http://root.cern.ch>

ALICE analysis modules

ALICE analysis modules

RAA\_1 | RAA\_2 | Stargenesis | Pt tutorial | Event display exercise | W0

W0

dataset: 1

OK

CernVM

MASTERCLASS MENU INSTRUCTIONS

Belongs to ALICE analysis modules!

Use the large picture button showing the ALICE logo to open the documentation for the event display and RAA analysis.

The picture buttons below will download, unpack and launch the corresponding analysis module. Use the /info buttons in each frame to get a description of what each module does.

You can quit at any time by clicking the 'quit' button at the menu. To restart the masterclass, open a terminal by clicking the terminal emulator icon on the bottom left. Then type:

```
[alice@localhost analysis] root masterclass.C
```



# CMS has more expansive Open Data Program for Research

We see external eco-system developing

- Workshop in October [link]

## Number of Papers appearing on e.g. Machine Learning methods for LHC

**EnergyFlow**

Search docs

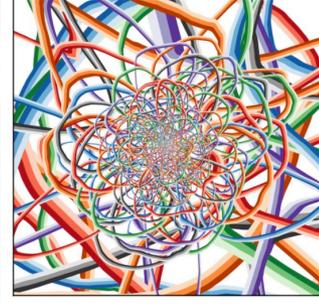
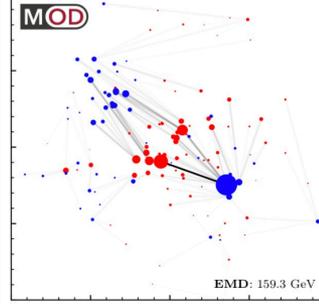
Home  
Getting Started  
Installation  
Demos  
Examples  
FAQs  
Release Notes  
News  
Documentation  
Architectures  
Datasets  
CMS Open Data  
HDF5 Format  
MOD  
GitHub

Docs > Documentation > Datasets

**CMS Open Data and the MOD HDF5 Format**

Starting in 2014, the CMS Collaboration began to release research-grade recorded and simulated datasets on the [CERN Open Data Portal](#). These fantastic resources provide a unique opportunity for researchers with diverse connections to experimental particle physics world to engage with cutting edge particle physics by developing tools and testing novel strategies on actual LHC data. Our goal in making portions of the CMS Open Data available in a reprocessed format is to ease as best as possible the technical complications that have thus far been present when attempting to use Open Data (see also [recent efforts by the CMS Collaboration](#) to make the data more accessible).

To facilitate access to Open Data, we have developed a format utilizing the widespread [HDF5 file format](#) that stores essential information for some particle physics analyses. This "MOD HDF5 Format" is currently optimized for studies based on jets, but may be [extended in the future to support other types of analyses](#).

EMD: 159.3 GeV

arxiv:1908.08542

arxiv:1910.07029

arxiv:1805.00850

**Exploring the Space of Jets with CMS Open Data**  
Patrick T. Komiske<sup>1,2,\*</sup>, Radha Mastandrea<sup>1,1</sup>, Eric M. Metodiev<sup>1,2,1</sup>, Preksha Naik<sup>1,3</sup> and Jesse Thaler<sup>1,2,4</sup>

<sup>1</sup>Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA  
<sup>2</sup>Department of Physics, Harvard University, Cambridge, MA 02138, USA

**End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data**  
M. Andrews<sup>1</sup>, J. Alison<sup>1</sup>, S. An<sup>1,2</sup>, P. Brvant<sup>1</sup>, B. Burkle<sup>3</sup>, S. Glevzer<sup>4</sup>, M. Narain<sup>5</sup>, M. Paulini<sup>1</sup>, B.

Noname manuscript No.  
(will be inserted by the editor)

<sup>1</sup>Department  
<sup>3</sup>Departme  
<sup>4</sup>Departme  
<sup>5</sup>Machine Learnin

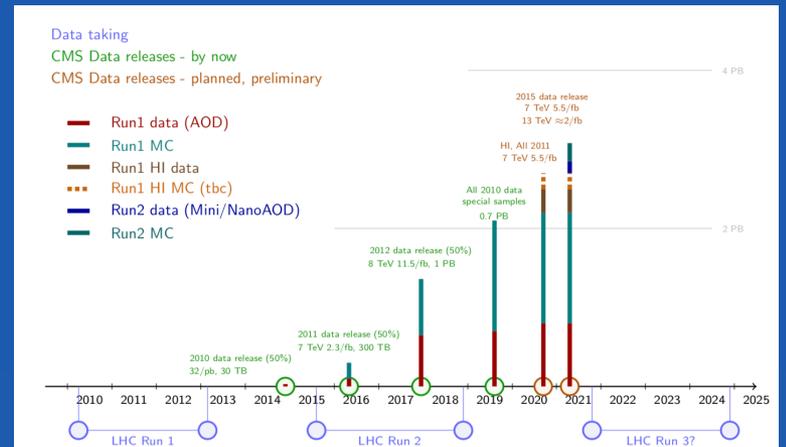
**Fast and accurate simulation of particle detectors using generative adversarial networks**  
Pasquale Musella · Francesco Pandolfi

26 Nov 2018  
the date of receipt and acceptance should be inserted later

**Abstract** Deep generative models parametrised by neural networks have recently started to provide accurate results in modeling natural images. In particular, generative adversarial networks provide an unsupervised solution to this problem. In this work we apply this

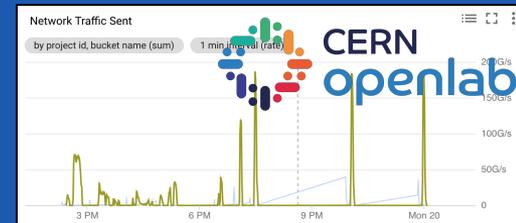
e.g. [2] and [3]) are based on estimators of particle trajectories and energy deposits. This information is subsequently aggregated in order to reconstruct energy, type and direction of final state particles produced by the collision of the primary beams.

Release Schedule being finalized for coming years



## Currently we see a range in approaches. Questions raised by Open Data (L3):

- Can we ensure ability to perform analysis of sufficient quality
- What data formats / software would be released? What's the level of support (if any?)
  - how "final" should objects be (ability to re-reconstruct...)
- Protect Collaboration / Cohesion
  - without a strong collaboration preparation of high-quality OD impossible
- Is analyzing PB scale data really feasible for external users.
  - emerging public cloud infrastructure might help provide on-demand access to scale
  - maybe targeted datasets (e.g. for ML R&D) rather than blanket Open Data?



KubeCon 2019 Keynote [link]

**CERN is seeking to harmonize them via a common Open Data Working Group.**

**HepData is for experimentalists interacting with wider community by releasing public information about already existing data analyses.**

**Open Data might be useful for experimentalists to interact with external researches on new R&D**

**But both approaches are not enough.**

**HepData is analysis-specific, but lossy.  
Open Data for new work outside of expt's**

**→ need infrastructure for internal, lossless analysis preservation**

Internal analysis preservation can capture detail unavailable in other modes of data/analysis preservation.

Increasing complexity in analyses to fully exploit potential of LHC dataset

- low-level observables: (e.g. BDT on calorimeter clusters)
  - even if we publish BDT / NN weights, can you reliably simulate those low-level details?
- whole-event observables (NN inputs from many objects)
  - simple description of signal model acceptance via e.g. efficiency tables will not work anymore
- Need a way to preserve analyses part of result pipeline at full fidelity.



## Internal Reuse:

Efforts by all LHC experiments to foster internal analysis preservation.  
Ingredients for AP:

### capture software

archive analysis code incl.  
dependencies

### capture commands

what do with the  
captured software

### capture workflow

order of individual steps

### data assets

input data needed  
to run the analysis

CERN provides infrastructure to  
assist experiments

REANA: workflows-as-a-service

CAP: store workflow and  
other analysis artifacts  
(software, etc)

**reana**

Reproducible research data analysis platform

CERN  
Analysis Preservation

capture, preserve and reuse physics analyses



## 1. capture software

archive analysis  
code incl. deps.

## 2. capture commands

what do with the  
captured software

## 3. capture workflow

order of individual  
steps

# Containers universally seen as suitable technology:

# all experiments have some infrastructure to run experiment / analysis code in containers

Two Docker Hub entries for ATLAS analysis containers:

- atlas/athanalysis**: By atlas • Updated 8 days ago. 1M+ Downloads, 3 Stars. ATLAS Athena Analysis Release. Container.
- atlas/analysisbase**: By atlas • Updated 8 days ago. 1M+ Downloads, 10 Stars. ATLAS Standalone Analysis Release. Container.

## REANA Environment AliPhysics

build unknown gliter join chat License GPL v2

### About

reana-env-aliphysics provides a container image with encapsulated runtime execution environment for AliPhysics based ALICE data analyses. The container image includes all the necessary dependencies and does not have any external requirements (such as CVMFS).

reana-env-aliphysics was developed for use in the REANA reusable research data analysis platform.

lhcb-analysis-preservation > containerization-cookie > Details

C

### containerization-cookie

Project ID: 31307 | Leave project

45 Commits 1 Branch 0 Tags 287 KB Files

Cookiecutter template for analysis containerization

## clelange / cmssw-docker

Code Issues 5 Pull requests 0 Actions Projects 0

Dockerfiles for CMSSW <https://doi.org/10.5281/zenodo.3374807>

82 commits 1 branch 0 packages

Tag: v1.0 New pull request

clelange Use --build-arg instead of wrong -e for docker ENV

## 1. capture software

archive analysis  
code incl. deps.

## 2. capture commands

what do with the  
captured software

## 3. capture workflow

order of individual  
steps

Workflow languages seem to be a good choice:

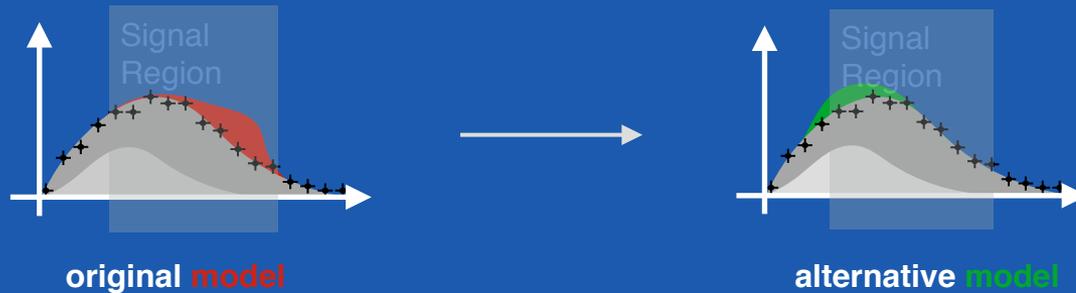
REANA supports Common Workflow Language, Yadage

• looking into snakemake (LHCb also has snakemake starter kit)

The image displays three overlapping screenshots of GitHub README pages for REANA analysis examples. The top-left screenshot shows the 'REANA example - ALICE LEGO train test run' with a 'build passing' badge and a 'license GPL-2.0' badge. The middle screenshot shows the 'REANA example - LHCb Rare Charm Decay Search' featuring a mathematical equation  $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$  and a 'license MIT' badge. The bottom-right screenshot shows the 'REANA example - CMS Higgs-to-four-leptons' with a 'license MIT' badge. Each page includes an 'About' section and an 'Analysis structure' section. The bottom-left screenshot shows a terminal window with the following commands:

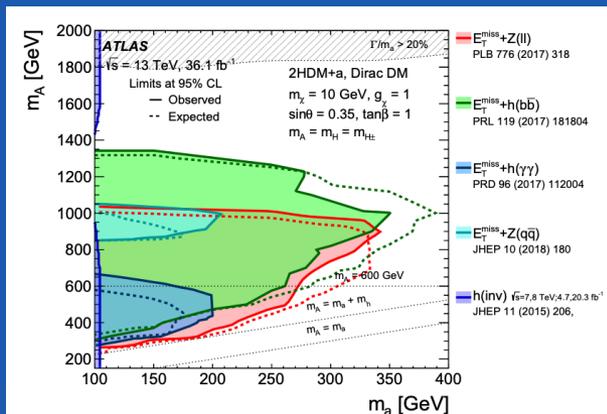
```
$ mkdir -p _alice_data_2010_LHC10h_2_000139038
$ cd _alice_data_2010_LHC10h_2_000139038
$ wget http://opendata.cern.ch/record/1102/files/assets/aLice/2010/LHC10h/000139038/ESD/000139038.root
$ cd ..
```

# Major use-case for internal re-use: reinterpretation

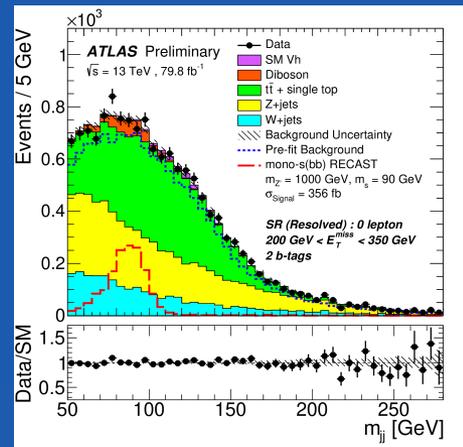


ATLAS: require analyzers to preserve analysis that at least reinterpretation w/ REANA is possible → realization of RECAST (docker images, scripts, workflows)

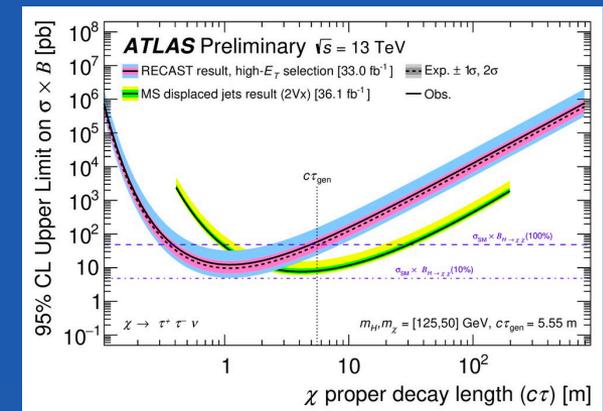
New scientific results based on this (rather technical) requirement



arxiv:1903.01400



ATL-PHYS-PUB-2019-032



ATL-PHYS-PUB-2020-007



# CERN Analysis Preservation Examples

LHCb

CMS

## Focus on:

- ease of use for analysis teams
  - e.g. auto-complete, automatic ingestion, command line clients
- ease of use for users
  - discoverability / search
  - integration with REANA



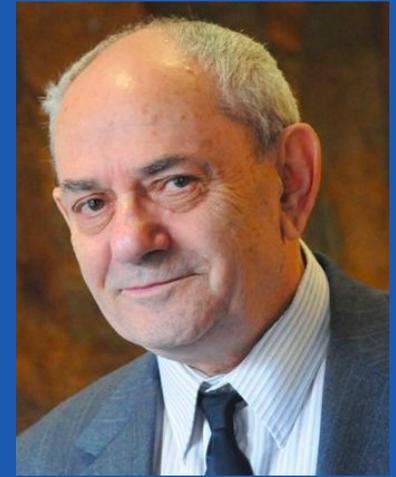
**We see a pattern:**

**Likelihood: Preserving general likelihood functions in a sustainable way is a hard problem.**

→ **restricting to binned models (HistFactory) enabled progress by narrowing scope**

**AP: Preserving Analyses in full generality (does it work in 100 years?) is too big a problem. RECAST focuses on**

- **near-/mid-term solution (e.g. assume containers)**
- **subset of the problem of reinterpretation, not e.g. re-estimation of background)**



"When solving a problem of interest, do not solve a more general problem as an intermediate step"  
- V. Vapnik

**We seem to be entering a golden era of data / analysis preservation for LHC.**

**Absence of new physics (so far!) forces us to focus on full exploitation of data.**

**We finally have the technology:**

- **containers / workflows for software/analysis preservation**
- **at-scale compute on-demand for Open Data**

**See some sociological shifts in community:**

- **likelihood releases**
- **analysis preservation (RECAST) as approval requirement**

**Good opportunity for new strategic efforts for both open and internal preservation efforts.**

